Розділ 8

Вибірковий метод

8.1. Поняття вибіркового спостереження та його теоретичні основи

Як зазначалось у розділі 3, за ступенем охоплення одиниць досліджуваної сукупності статистичне спостереження може бути суцільним і несуцільним.

Суцільне спостереження передбачає обстеження усіх без винятку одиниць генеральної сукупності. Наприклад, для визначення загальної чисельності населення під час перепису збирають дані про кожну окрему людину, яка проживає в країні, для встановлення обсягу виробленої продукції (сталі, вугілля, молока, м'яса тощо), ведуть щоденний облік її виходу і т. д.

Проте з деяких причин (великої трудомісткості, тривалості проведення, високої вартості тощо) суцільне спостереження часто буває економічно недоцільним або практично нездійсненним. Тому на практиці переважно застосовують несуцільне спостереження, різновидом якого є вибіркове.

Вибірковим спостереженням в статистиці називають такий вид спостереження, який дає можливість зробити висновок про всю сукупність одиниць при обстеженні тільки її частини. Прикладом вибіркового спостереження може бути вибіркове обстеження домогосподарств, визначення втрат урожаю, якості продукції, польові досліди тощо.

Сукупність методів математичної статистики, що застосовуються для обґрунтувань та висновків при проведенні вибіркового спостереження, називають **вибірковим методом.**

Вибіркове спостереження є одним з видів несуцільного спостереження, що найбільш широко застосовується. Вибірковими даними користуються досить широко в різних сферах суспільного життя. Наприклад, для оцінки якості продукції (зерна, молока тощо) немає необхідності в дослідженні всього обсягу продукції, для цього досить лише взяти певну кількість проб.

Вибіркове дослідження одержало широке поширення в державній і відомчій статистиці. В статистичній практиці вибіркове спостереження застосовується для обстеження домогосподарств населення, житлових умов сімей робітників і службовців, їх заробітної плати, цін на ринках, в науковій роботі при статистичній обробці результатів дослідження, для вивчення і контролю якості продукції, громадської думки тощо.

Із розділів 3 і 4 відомо, що основною формою збирання кримінологічної і соціально-правової інформації є державна і відомча статистична звітність правоохоронних органів та інших юридичних установ. Статистична звітність відбиваючи найважливіші показники роботи правоохоронних органів разом з тим обмежена обсягом і тому не може охопити своїми даними всю багатогранну діяльність цих органів. У зв'язку з цим по актуальних питаннях, не відображених в офіційній звітності або відображеній з недостатньою повнотою, слід проводити спеціальні вибіркові спостереження правових явищ і процесів з подальшим використанням одержаної інформації для аналізу рівня, структури і динаміки правопорушень і стану криміногенної ситуації в країні і окремих її регіонах, а також для уточнення і розробки даних суцільного обстеження.

Джерелами інформації при організації і проведенні вибіркового спостереження в правовій статистиці є дані кримінальних і цивільних справ, справ про адміністративні правопорушення, статистичної звітності, документів первинного обліку (статистичних карток, журналів, довідок тощо), інформаційних бюлетенів, оглядів, аналітичних довідок, заяв громадян і посадових осіб, та інші документи, які містять відомості про злочин і злочинця, потерпілі, члени родини злочинця, інші родичі, друзі, оточення за місцем роботи і проживання тощо.

Особливе місце вибіркове спостереження займає в дослідженнях злочинності і пов'язаних з нею проблем. У кримінально-правовій статистиці прикладом вибіркового спостереження є вивчення середнього віку засуджених, середніх строків позбавлення волі, питомої ваги засуджених неповнолітніх серед засуджених за певний вид злочину, питомої ваги злочинів вчинених у стані сп'яніння, відмінених вироків тощо.

При дотриманні правил наукової організації обстеження вибіркове спостереження дає досить точні результати, тому його часто застосовують для уточнення даних суцільного обліку. Так, при проведенні перепису населення організовують вибіркові контрольні обходи для перевірки правильності записів суцільного спостереження, під час перепису (обліку) худоби на основі вибіркових даних визначають процент недообліку худоби у населення та ін.

В деяких випадках вибіркове спостереження використовують для дослідження процесів і явищ, суцільне спостереження яких недоцільне внаслідок занадто великого обсягу робіт, наприклад обстеження домогосподарств, реєстрація цін на ринках, вивчення споживчого попиту населення тощо.

У проведенні ряду досліджень вибіркове спостереження є єдино можливим, наприклад при контролі якості продукції, якщо перевірка супроводжується знищенням або псуванням зразків, що обстежуються (визначення жирності молока, схожості зерна, міцності тканин на розрив, тривалості горіння електричних ламп тощо).

Широкому використанню вибіркового спостереження в статистичній практиці сприяє сучасна обчислювальна техніка, яка дає змогу значно ускладнити процедури обробки вихідної інформації, а тому підвищити надійність статистичних показників, що одержують за даними вибірки.

З усіх видів несуцільного спостереження головним є вибіркове спостереження, оскільки тільки дані цього спостереження дають змогу науково обґрунтовано і з необхідною вірогідністю поширити дані, одержані по частині сукупності на всю сукупність.

Науково організоване вибіркове спостереження має ряд суттєвих переваг перед суцільним.

1. Економія трудових, матеріальних ресурсів і коштів за рахунок скорочення обсягів робіт по збиранню, обробці і узагальненню даних. Наприклад, якщо вибірці підлягає 10% загальної чисельності одиниць, то обсяг робіт скорочується порівняно з суцільним обстеженням в 10 разів.

2. Проведення спостереження у стислі строки і за більш широкою програмою, одержання кінцевих результатів дослідження в короткі строки.

3. Зведення до мінімуму псування або навіть знищення досліджуваних зразків при перевірці їх якості.

4. Досягнення більшої точності результатів спостереження завдяки скороченню помилок, що виникають при реєстрації.

В процесі проведення вибіркового спостереження необхідно дати правильне уявлення про зведені показники всієї сукупності на основі обстеження її частини при умові дотримання всіх вимог і принципів проведення статистичного спостереження і науково організованої роботи по відбору одиниць.

Сукупність відібраних для обстеження одиниць прийнято називати **вибірковою,** а сукупність одиниць, з якої проводиться відбір, — **генеральною.**

Властивість вибіркової сукупності відтворювати генеральну сукупність дістала назву **репрезентативності,** що означає представництво з певною точністю і вірогідністю. В зв'язку з тим, що при вибірковому спостереженні обстежується тільки частина одиниць генеральної сукупності, то характеристики вибіркової сукупності, як правило, відрізняються від характеристик вибіркової сукупності, тобто мають місце так звані **помилки репрезентативності** (відповідності, відображення).

Тому одним із основних завдань вибіркового методу є одержання таких вибіркових характеристик, які б якомога точніше відтворювали характеристики генеральної сукупності, тобто давали найменші помилки репрезентативності.

Об'єктивну гарантію репрезентативності одержаної вибіркової сукупності дає застосування відповідних науково обґрунтованих способів відбору одиниць, що підлягають обстеженню. В процесі формування вибіркової сукупності має бути забезпечений строго об'єктивний підхід до відбору одиниць, що гарантує рівну можливість кожній одиниці генеральної сукупності потрапити у вибірку. Кількість відібраних при цьому одиниць має бути досить великою.

Мета вибіркового спостереження полягає в тому, що на основі відібраних з генеральної сукупності для обстеження одиниць необхідно дати оцінку невідомим параметрам генеральної сукупності.

Теоретичною основою вибіркового методу є математичні теореми закону великих чисел, викладення яких дається в курсі математичної статистики, і теорія імовірностей.

На основі теорем закону великих чисел розв'язуються два взаємопов'язаних і важливих в практичному відношенні питання вибіркового спостереження: розрахунок необхідної чисельності вибірки і визначення помилок вибірки при заданому рівні довірчої імовірності.

Результати вибіркового спостереження характеризуються середніми і відносними узагальнюючими показниками. Узагальнюючі показники генеральної сукупності (середня, частка, дисперсія та ін.) називають **генеральними**, а відповідні узагальнюючі показники вибіркової сукупності - **вибірковими.**

У вибірковому спостереженні прийняті такі позначення. Обсяг генеральної сукупності позначають через ***N***, а вибіркової - через ***n***. Середню величину і дисперсію ознаки в генеральній сукупності називають **генеральною середньою** і генеральною **дисперсією.** Генеральну середню позначають через $\bar{x}$, а генеральну дисперсію – через σ_0^2.

Середню величину і дисперсію ознаки у вибірковій сукупності називають **вибірковою середньою** і **вибірковою дисперсією**. Вибіркову середню позначають через $\tilde{x}$, а вибіркову дисперсію – через σ^2.

Частку одиниць, що володіють даною ознакою в генеральній сукупності, називають **генеральною часткою**. Відносна величина частки, що одержана в результаті вибіркового спостереження, називається **вибірковою часткою** або **частістю**. Частість показує, яка частка одиниць вибіркової сукупності володіє досліджуваною ознакою. Генеральну частку позначають через ***p***, а частість - через ***ω***.

Наведемо приклад розрахунку зазначених показників для генеральної і вибіркової сукупності. Припустимо, що з 300 ув'язнених віком 24 – 26 рооків, що представляють генеральну сукупність, відібрано у випадковому порядку 30 ув'язнених (10% усіх ув'язнених), що становлять вибіркову сукупність (табл. 8.1).

Таблиця 8.1

Вік ув'язнених і чисельність генеральної і вибіркової сукупностей

Групи ув'язнених за віком, років *x*	Чисельність засуджених	
	усього (генеральна сукупність) *N*	в тому числі відібрано (вибіркова сукупність) *n*
24	70	6
25	150	15
26	80	9
Разом	**300**	**30**

Визначимо за цими даними середній вік, дисперсію і частку ув'язнених віком 25 років і більше для генеральної і вибіркової сукупностей:

а) **для генеральної сукупності**

середній вік ув'язнених

$$x = \frac{\Sigma xf}{\Sigma f} = \frac{24 \cdot 70 + 25 \cdot 150 + 26 \cdot 80}{70 + 150 + 80} = \frac{7510}{300} = 25{,}0 \text{ років}$$

дисперсія

$$\sigma_0^2 = \frac{\Sigma(x - \bar{x})^2 f}{\Sigma f} = \frac{(24-25)^2 70 + (25-25)^2 150 + (26-25)^2 80}{70 + 150 + 80} = \frac{150}{300} = 0{,}50;$$

частка ув'язнених віком 25 років і більше

$$p = \frac{230}{300} = 0{,}77, \quad \textit{або} \quad 77\%;$$

б) **для вибіркової сукупності**

середній вік ув'язнених

$$\tilde{x} = \frac{\Sigma xf}{\Sigma f} = \frac{24 \cdot 6 + 25 \cdot 15 + 26 \cdot 9}{6 + 15 + 9} = \frac{753}{30} = 25{,}1 \text{ року};$$

дисперсія

$$\sigma^2 = \frac{\Sigma(x - \tilde{x})^2 f}{\Sigma f} = \frac{(24 - 25{,}1)^2 6 + (25 - 25{,}1)^2 15 + (26 - 25{,}1)^2 9}{6 + 15 + 9} = \frac{14{,}70}{30} = 0{,}49;$$

частка ув'язнених віком 25 років і більше

$$\omega = \frac{24}{30} = 0{,}80, \quad \textit{або} \quad 80\%.$$

Отже, генеральна середня становить 25,0 років, генеральна дисперсія - 0,50, а генеральна частка - 0,77, відповідно вибіркова середня - 25,1 року, вибіркова дисперсія - 0,49. вибіркова частка - 0,80, Як видно, ті самі показники генеральної і вибіркової сукупностей не збігаються. В абсолютному виразі різниця між середніми становить 0,1 року, між частками - 0,03. Незначні розбіжності між середніми і відносними показниками свідчать проте, що вибіркова сукупність досить точно репрезентує генеральну сукупність.

8.2. Помилки вибірки

Між показниками вибіркової сукупності і шуканими показниками (параметрами) генеральної сукупності, як правило, існують деякі розбіжності, які називають **помилками вибірки**. Загальна помилка вибіркової характеристики складається з помилок двох родів: помилок реєстрації і помилок репрезентативності.

Помилки реєстрації властиві будь-якому статистичному спостереженню і поява їх може бути викликана неуважністю реєстратора, неточністю підрахунків, недосконалістю вимірювальних приладів тощо.

Помилки репрезентативності притаманні тільки вибірковому спостереженню і зумовлені самою його природою оскільки як би ретельно і правильно не проводився відбір одиниць середні і відносні показники вибіркової сукупності завжди будуть певною мірою відрізнятися від відповідних показників генеральної сукупності.

Розрізняють систематичні та випадкові помилки репрезентативності. **Систематичні помилки репрезентативності** — це неточності, які виникають внаслідок недотримання умов відбору одиниць у вибіркову сукупність, не надання рівної можливості кожній одиниці генеральної сукупності потрапити у вибірку. **Випадкові помилки репрезентативності** — це похибки, які виникають внаслідок того, що вибіркова сукупність точно не відтворює характеристики генеральної сукупності (середню, частку, дисперсію та ін.) в силу несуцільного характеру обстеження.

При дотриманні принципу випадкового відбору розмір помилки вибірки залежить насамперед від чисельності вибірки. Чим більше чисельність вибірки при інших рівних умовах, тим меншою є величина помилки вибірки. При великій чисельності вибірки виразніше проявляється дія закону великих чисел, згідно з яким: з імовірністю, скільки завгодно близькою до одиниці, можна стверджувати, що при досить великому обсязі вибірки та обмеженій дисперсії вибіркові характеристики (середня, частка) будуть скільки завгодно мало відрізнятися від відповідних генеральних характеристик.

Розміри помилки вибірки також безпосередньо пов'язані зі ступенем варіювання досліджуваної ознаки, а ступінь варіювання, як зазначалося вище, в статистиці характеризується розміром дисперсії (розсіяння): чим менша дисперсія, тим меншою є помилка вибірки, тим надійніші статистичні висновки. Тому на практиці дисперсію ототожнюють з помилкою вибірки.

Оскільки параметр генеральної сукупності є шукана величина і він невідомий, потрібно орієнтуватися не на конкретну помилку, а середню з усіх можливих вибірок.

Якщо з генеральної сукупності відібрати кілька вибіркових сукупностей, то кожна із отриманих вибірок дасть різне значення конкретної помилки.

Середня квадратична величина μ, обчислена з усіх можливих значень конкретних помилок(ε_i) становитиме:

$$\mu = \sqrt{\frac{\Sigma \varepsilon_i^2 f_i}{\Sigma f_i}} = \sqrt{\frac{\Sigma(\tilde{x}_i - \bar{x})^2 f_i}{\Sigma f_i}},$$

де $\tilde{x}_i$ — вибіркові середні; $\bar{x}$ — генеральна середня; f_i— чисельність вибірок з величиною $\varepsilon_i = \tilde{x}_i - \bar{x}$.

Середнє квадратичне відхилення вибіркових середніх від генеральної середньої називають **середньою помилкою вибірки.**

Залежність величини помилки вибірки від її чисельності та від ступеня варіювання ознаки знаходить вираження у формулі середньої помилки вибірки μ.

Квадрат середньої помилки (дисперсія вибіркових середніх) прямо пропорційній дисперсії σ_0^2 і обернено пропорційній чисельності вибірки *n*:

$$\mu^2 = \frac{\sigma_0^2}{n},$$

де σ_0^2 - дисперсія ознаки у генеральній сукупності.

Звідси середню помилку в загальному вигляді визначають за формулою:

$$\mu = \frac{\sigma_0}{\sqrt{n}}.$$

Отже, визначивши за вибіркою середнє квадратичне відхилення, можна встановити значення середньої помилки вибірки, величина якої, як випливає з формули, тим більша, чим більшою є варіація випадкової величини і тим менша, чим більшою є чисельність вибірки.

Тому по мірі зростання обсягу вибірки розмір середньої помилки зменшується. Якщо, наприклад, потрібно зменшити середню помилку вибірки в два рази, то чисельність вибірки слід збільшити в чотири рази, якщо треба зменшити помилку вибірки в три рази, то обсяг вибірки слід збільшити в дев'ять разів і т. д.

У практичних розрахунках застосовують дві формули середньої помилки вибірки: для середньої і для частки.

При вибірковому вивченні середніх показників формула середньої помилки така:

$$\mu_{\tilde{x}} = \sqrt{\frac{\sigma_0^2}{n}}.$$

При вивченні відносних показників (часток ознак) формула середньої помилки має вигляд:

$$\mu_p = \sqrt{\frac{p(1-p)}{n}},$$

де *p* — частка ознаки в генеральній сукупності.

Застосування наведених формул середньої помилки передбачає, що відомі генеральна дисперсія та генеральна частка. Проте в дійсності ці показники невідомі і обчислити їх неможливо через відсутність даних щодо генеральної сукупності. Тому виникає потреба заміни генеральної дисперсії та генеральної частки іншими, близькими до них, величинами.

В математичній статистиці доведено, що такими величинами можуть бути вибіркова дисперсія(σ^2) та вибіркова частка ω.

З урахуванням сказаного формули середньої помилки можуть бути записані так:

$$\mu_{\tilde{x}} = \sqrt{\frac{\sigma^2}{n}}; \qquad \mu_{\omega} = \sqrt{\frac{\omega(1-\omega)}{n}}.$$

Ці формули дають змогу визначити середню помилку при повторній вибірці. Застосування простої випадкової повторної вибірки у практиці є обмеженим. Насамперед практично недоцільно, а інколи неможливо повторне обстеження тих самих одиниць. Застосування безповторного відбору замість повторного диктується також вимогою підвищення ступеня точності і надійності вибірки. Тому на практиці найчастіше використовують спосіб безповторного випадкового відбору. За цим способом відбору одиниця сукупності, що відібрана у вибірку, в подальшому відборі участі не бере. Одиниці відбирають із генеральної сукупності, зменшеної на кількість раніше відібраних одиниць. Тому в зв'язку із зміною чисельності генеральної сукупності після кожного відбору та ймовірності відбору для одиниць, що залишились, у формули середньої помилки вибірки вводиться поправочний множник

$$\frac{N-n}{N-1},$$

де *N* - чисельність генеральної сукупності; *n* - чисельність вибірки. При досить великому значенні *N* можна одиницею в знаменнику знехтувати. Тоді

$$\frac{N-n}{N-1} \approx \frac{N-n}{N} \approx 1 - \frac{n}{N}.$$

Відтак формули середньої помилки вибірки для безповторного відбору для середньої і для частки відповідно мають вигляд:

$$\mu_{\tilde{x}} = \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}; \qquad \mu_\omega = \sqrt{\frac{\omega(1-\omega)}{n}\left(1-\frac{n}{N}\right)}.$$

Оскільки n завжди менше N, то додатковий множник $1-\frac{n}{N}$ завжди менше одиниці. Отже, абсолютне значення помилки вибірки при безповторному відборі завжди менше, чим при повторному.

Якщо чисельність вибірки досить велика, то величина $1-\frac{n}{N}$ близька до одиниці, а тому нею можна знехтувати. Тоді середню помилку випадкового безповторного відбору визначають за формулою власне випадкової повторної вибірки.

Розрахуємо для нашого прикладу середню помилку для віку і для частки ув'язнених віком 25 років і більше.

Середня помилка вибірки

а) середнього віку ув'язнених

$$\mu_{\tilde{x}} = \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)} = \sqrt{\frac{0{,}50}{30}\left(1-\frac{30}{300}\right)} = \sqrt{0{,}015} = \pm 0{,}12 \text{ року};$$

б) частки ув'язнених віком 25 років і більше

$$\mu_p = \sqrt{\frac{p(1-p)}{n}\left(1-\frac{n}{N}\right)} = \sqrt{\frac{0{,}77 \cdot 0{,}23}{30}\left(1-\frac{30}{300}\right)} = \sqrt{0{,}0053} = \pm 0{,}07.$$

Середній вік ув'язнених в генеральній сукупності $\tilde{x} \pm \mu_{\tilde{x}} = 25{,}1 \pm 0{,}12$ років, тобто знаходиться в межах від 24,98 до 25,22 року.

Частка ув'язнених віком 25 років і більше в генеральній сукупності $p = \omega \pm \mu_p = 0{,}80 \pm 0{,}07$, тобто знаходиться в межах від 73 до 87%.

Середня помилка вибірки показує можливі відхилення характеристик вибіркової сукупності від характеристик генеральної сукупності. Разом з тим при проведенні вибіркового спостереження перед дослідниками часто стоїть завдання розрахунку не тільки середньої помилки, але і визначення граничної можливої помилки вибірки. Знаючи середню помилку, можна визначити межі, за які не вийде величина помилки вибірки. Однак стверджувати, що ці відхилення не перевищать заданої величини, можна не з абсолютною вірогідністю, а лише з певним ступенем імовірності. Рівень імовірності, що приймається при визначенні можливих меж, в яких містяться значення параметрів генеральної сукупності, називається **довірчим рівнем імовірності.**

Довірча імовірність – це досить висока і, така, що практично вважається здійсненою в кожному конкретному випадку, імовірність, що гарантує отримання надійних статистичних висновків. Позначимо її через P, а імовірність перевищити цей рівень – α. Отже, $\alpha = 1 - P$. Імовірність α **називають рівнем значущості** (істотності), який характеризує відносне число помилкових висновків у загальному числі висновків і визначається як різниця між одиницею і довірчою імовірністю, що приймається.

Рівень довірчої імовірності встановлює дослідник виходячи зі ступеня відповідальності і характеру завдань, що розв'язуються. У статистичних дослідженнях суспільних явищ найчастіше приймається рівень довірчої імовірності $P = 0{,}95$; $P = 0{,}99$ (відповідно рівень значущості $\alpha = 0{,}05$; $\alpha = 0{,}01$) рідше $P = 0{,}999$. Наприклад, довірча імовірність $P = 0{,}99$ означає, що помилка оцінки у 99 випадках із 100 не перевищить встановленої величини і тільки в одному випадку із 100 може досягти обчисленого значення, або перевищити його.

Помилка вибірки, що обчислена із заданим ступенем надійної імовірності, називається **граничною помилкою вибірки ε_p**.

Розглянемо, як встановлюється величина можливої граничної помилки вибірки. Величина ε_p пов'язана з нормованим відхиленням t, яке визначається як відношення граничної помилки вибірки ε_p до середньої помилки μ:

$$t = \frac{\varepsilon_p}{\mu}.$$

Для зручності розрахунків відхилення випадкової величини від її середнього значення звичайно виражають в одиницях середнього квадратичного відхилення. Вираз

$$\frac{x_i - \bar{x}}{\sigma}$$

називають **нормованим відхиленням.** В статистичній літературі t ще називають **коефіцієнтом довіри,** або **коефіцієнтом кратності середньої помилки вибірки.**

Так, нормоване відхилення для вибіркової середньої можна визначити за формулою:

$$t = \frac{\varepsilon_p}{\mu_{\tilde{x}}} = \frac{\tilde{x} - \bar{x}}{\mu_{\tilde{x}}} = \frac{\tilde{x} - \bar{x}}{\sigma}\sqrt{n},$$

де $\mu_{\tilde{x}} = \frac{\sigma}{\sqrt{x}}$.

Із виразу $t = \dfrac{\varepsilon_p}{\mu}$ можна знайти можливу граничну помилку, вибірки $\varepsilon_p = t\mu$.

Підставивши замість μ її значення, наведемо формули граничних помилок вибірки для середньої і для частки при безповторному випадковому відборі:

$$\varepsilon_{\tilde{x}} = t\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}; \qquad \varepsilon_p = t\sqrt{\frac{\omega(1-\omega)}{n}\left(1-\frac{n}{N}\right)}.$$

Отже, гранична помилка вибірки залежить від величини середньої помилки і нормованого відхилення і дорівнює ± кратному числу середніх помилок вибірки.

Середня і гранична помилки вибірки – величини іменовані і виражаються в тих самих одиницях, що й середня арифметична і середнє квадратичне відхилення.

Нормоване відхилення функціонально зв'язано з імовірністю.

Імовірність відхилень вибіркових середніх від генеральної середньої ($\tilde{x} - \bar{x}$) при великому числі спостережень ($n \to \infty$) визначається законом нормального розподілу Лапласа-Гаусса.

Нормальним розподілом називають розподіл неперервної випадкової величини, який описується щільністю імовірності

$$\Phi(x) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{\frac{(x_i-\bar{x})^2}{2\sigma^2}},$$

де $\Phi(x)$ – щільність імовірності (ордината кривої); σ_0 – середнє квадратичне відхилення генеральної середньої, яке у практичних розрахунках замінюється вибірковим σ; π = 3,14 ... (постійна величина, яка характеризує відношення довжини кола до довжини його діаметра); e = 2,718... – основа натуральних логарифмів (число Ейлера).

Як видно, нормальний розподіл визначається двома параметрами: середньою арифметичною і середнім квадратичним відхиленням. Знаючи ці параметри, можна побудувати криву нормального розподілу.

Звичайно $\dfrac{x_i - \bar{x}}{\sigma}$ в даній формулі замінюється на t, де відхилення представлені в частках середнього квадратичного відхилення, прирівняного до одиниці. Завдяки нормуванню, дисперсія $t = 1$, а $\bar{x} = 0$.

Рівняння нормальної кривої при такій заміні приймає вигляд:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Його називають **стандартним рівнянням нормальної кривої**, а нормальну криву – **нормованою кривою.**

Графік щільності нормального розподілу називається **нормальною кривою** або кривою Гаусса (рис. 8.1).

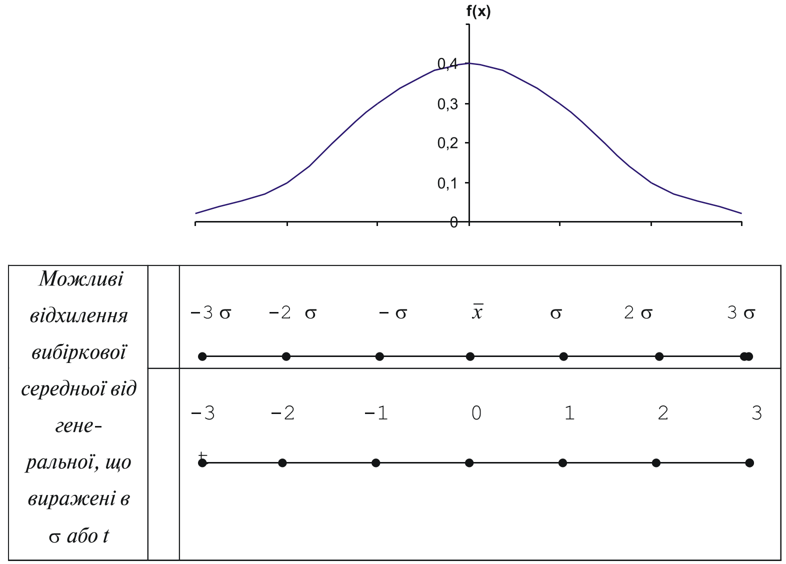

Рис. 8.1. Крива нормального розподілу імовірностей

Ця крива симетрична відносно осі ординат і асимптотично наближається до осі абсцис. Крива має точки перегину при $t = \pm 1$, тобто при таких відхиленнях значень ознаки від середньої арифметичної, які дорівнюють одному середньому квадратичному відхиленню. Площа, що обмежена кривою і віссю абсцис, дорівнює одиниці. Значення щільності імовірності $\Phi(x)$ залежить тільки від величини нормованого відхилення t, так як π і e – постійні величини. Так, при $t = 0$ співмножник $e^{-\frac{t^2}{2}} = 1$ і щільність імовірності максимальна $\Phi(0) = 0{,}3989$. По мірі зростання t щільність імовірності зменшується (дод. 15).

Для знаходження значень t складені спеціальні таблиці (дод.16), за якими можна знайти значення t при заданому рівні довірчої імовірності і значення імовірності при відомому t.

Наведемо значення *t* та відповідні до них імовірності для вибірок з чисельністю $n \geq 30$, що найчастіше використовується в практичних розрахунках:

t	1,00	1,96	2,00	2,58	3,00
P	0,6827	0,9500	0,9545	0,9901	0,9973

Отже, при $t = 1$ імовірність відхилення вибіркових характеристик від генеральних на величину однократної середньої помилки вибірки дорівнює 0,6827. Це означає, що в середньому із кожної 1000 вибірок 683 дадуть узагальнені характеристики, які відрізнятимуться від генеральних узагальнених характеристик не більше, чим на величину однократної середньої помилки. При $t = 2$ імовірність дорівнює 0,9545. Це означає, що із кожної 1000 вибірок 954 дадуть узагальнені характеристики, які відрізнятимуться від генеральних узагальнених характеристик не більш ніж на двократну середню помилку вибірки і т.д.

Однак в зв'язку з тим, що, як правило, проводиться тільки одна вибірка, то ми кажемо, що, наприклад, із імовірністю 0,9545 можна гарантувати, що розміри граничної помилки не перевищать двократну середню помилку вибірки.

Математично доведено, що відношення помилки вибірки до середньої помилки, як правило, не перевищує $\pm 3\mu$ при досить великій чисельності *n*, незважаючи на те, що помилка вибірки може набувати будь-які значення. Іншими словами можна сказати, що при досить високій імовірності судження ($P = 0,9973$) гранична помилка вибірки, як правило, не перевищує трьох середніх помилок вибірки. Тому величину $\varepsilon_p = 3\mu$ можна прийняти за межу можливої помилки вибірки.

Визначимо для нашого прикладу граничну помилку вибірки для середнього віку і для частки ув'язнених віком 25 років і більше. Довірчий рівень імовірності приймемо рівним $P = 0,9545$. За таблицею (дод.16) знайдемо значення $t = 2$. Середні помилки вибірки для віку і частки ув'язнених віком 25 років і більше були знайдені раніше і відповідно становили: $\mu_{\tilde{x}} = \pm 0,12$ року; $\mu_p = \pm 0,07$.

Гранична помилка середнього віку ув'язнених:

$$\varepsilon_{\tilde{x}} = t\mu_{\tilde{x}} = 2 \cdot 0,12 = \pm 0,24 \text{ року}.$$

Отже, різниця між вибірковим середнім віком і генеральною середньою буде не більше 0,24 року. Межі середнього віку ув'язнених в генеральній сукупності: $\bar{x} = \tilde{x} \pm \varepsilon_{\tilde{x}} = 25,1 + 0,24$, тобто від 24,86 до 25,34 року.

Гранична помилка частки ув'язнених 25 років і більше:

$$\varepsilon_p = t\mu_p = 2 \cdot 0,07 = \pm 0,14.$$

Отже, гранична помилка у визначенні частки ув'язнених віком 25 років і більше не перевищить 14%, тобто питома вага ув'язнених із зазначеним віком в генеральній сукупності знаходиться в межах: $p = \omega \pm \varepsilon_p = 0,80 \pm 0,14$, тобто від 66 до 94%.

Уникнути складних математичних розрахунків граничних помилок вибірки для **якісних ознак** при заданій чисельності вибірки можна за допомогою використання спеціальних таблиць з уже готовими результатами. Наведемо витяг з однієї таких таблиць (табл. 8.2).

Таблиця 8.2

Межа помилки при заданій чисельності вибірки t = 2,%[1]

Питома вага спостережень, %	Чисельність вибірки								
	100	200	300	400	500	600	700	800	900
10	6,0	4,3	3,5	3,0	2,7	2,5	2,3	2,1	2,0
15	7,2	5,1	4,1	3,6	3,2	2,9	2,7	2,5	2,4
20	8,0	5,7	4,6	4,0	3,6	3,3	3,0	2,8	2,7
30	9,2	6,5	5,3	4,6	4,1	3,7	3,5	3,2	3,1
35	9,6	6,8	5,5	4,8	4,3	3,9	3,6	3,4	3,2
40	9,9	7,0	5,6	4,9	4,4	4,0	3,7	3,5	3,3
45	10,0	7,1	5,7	5,0	4,5	4,1	3,8	3,5	3,3
55	10,0	7,1	5,7	5,0	4,5	4,1	3,8	3,5	3,3
65	9,6	6,8	5,5	4,8	4,3	3,9	3,6	3,4	3,2
70	9,2	6,5	5,3	4,6	4,1	3,7	3,5	3,2	3,1
75	8,7	6,2	5,0	4,3	3,9	3,5	3,3	3,1	2,9
80	8,0	5,7	4,6	4,0	3,6	3,3	3,0	2,8	2,7

Покажемо на прикладі, як треба нею користуватися. Припустимо, що на основі обстеження кримінальних справ 200 засуджених за хуліганство ми встановили, що 70% з них вчинили злочин у стані алкогольного сп'яніння. Граничну помилку знайдемо на перетині горизонтального рядка з числом 70 і вертикального стовпця з числом 200. При рівні ймовірності $P = 0,9545$ і коефіцієнти довіри $t = 2$ гранична помилка вибірки становитиме 6,5%. Це означає, що при даній чисельності вибірки (200 осіб) частка засуджених, вчинивших хуліганство в стані алкогольного сп'яніння в генеральній сукупності, може коливатися в межах від 63,5 до 76,5 % (70 ± 6,5).

[1] Боярский А.Я. Таблицы для определения достоверности статистических показателей и числа наблюдений в статистическом исследовании. М., 1947.

Оскільки всі елементи генеральної сукупності для обчислення шуканого параметра, як правило, використати неможливо, то про цей параметр намагаються судити за даними однієї або кількох вибірок із генеральної сукупності.

Наближене значення шуканого параметра генеральної сукупності, встановлене за даними вибіркової сукупності, називають **вибірковою оцінкою параметра.**

Оцінка невідомого параметра генеральної сукупності може бути проведена двояко: або одним числом (точною) - **точкова оцінка,** або із зазначенням інтервалу, в якому із заданою ймовірністю може знаходитись шуканий параметр, - **інтервальна оцінка**.

Суть точкової оцінки полягає в тому, що за найкращу оцінку шуканого параметру генеральної сукупності приймається знайдене за вибіркою його конкретне числове значення.

Результати точкової оцінки шуканого параметру генеральної сукупності можна записати таким чином: за статистичну оцінку параметра генеральної сукупності приймається його вибіркове значення з середньою помилкою $\pm \mu$.

Прикладом точкової оцінки параметра генеральної сукупності був розрахунок середнього віку ув'язнених за вибірковими даними ($\tilde{x} \pm \mu_{\tilde{x}} = 25{,}1 \pm 0{,}12$ року).

При невеликому обсязі вибірки точкова оцінка в значній мірі випадкова і малоефективна і тому може істотно відрізнятися від параметра генеральної сукупності, тобто приводити до великих відхилень вибіркових характеристик від генеральних. З цієї причини при невеликому обсязі вибірки доцільно користуватися інтервальною оцінкою.

Інтервальною називають оцінку, яка визначається двома числами - кінцями інтервалу, в якому із заданою імовірністю знаходиться шуканий параметр. Центром такого інтервалу, як правило, беруть знайдену вибіркову оцінку точки, а визначення самих кінців інтервалу пов'язується з середньою помилкою оцінки і довірчою імовірністю. Отже, інтервальна оцінка є подальшим доповненням і розширенням точкової оцінки.

Встановивши довірчу імовірність, можна побудувати довірчий інтервал. **Довірчим інтервалом** для параметра генеральної сукупності називається такий інтервал, відносно якого можна із заздалегідь встановленою довірчою імовірністю $P=1-\alpha$, близькою до одиниці, стверджувати, що він містить невідоме значення параметра генеральної сукупності. Іншими словами, це інтервал, який покриває невідомий параметр генеральної сукупності із заданою імовірністю P.

Наприклад, шуканий довірчий інтервал для оцінки генеральної середньої матиме вигляд:

$$I_p(\tilde{x} - t\mu_{\tilde{x}};\ \tilde{x} + t\mu_{\tilde{x}}),$$

де $t\mu = \varepsilon_p$ — гранична помилка оцінки.

Проведемо точкову і інтервальну оцінку генеральної середньої за даними великої вибірки на такому прикладі. Є дані щодо стажу роботи 30 слідчих: 2; 5; 15; 7; 18; 20; 9; 6; 18; 15; 4; 16; 25; 8; 30; 1; 26; 20; 21; 6; 35; 30; 18; 26; 31; 3; 24; 32; 17; 22. Довірчий рівень імовірності P = 0,9545, якому відповідає $t = 2$.

Для визначення середньої помилки вибірки визначимо середню арифметичну і дисперсію.

Середній стаж роботи слідчих

$$\bar{x} = \frac{\Sigma x}{n} = \frac{510}{30} = 17 \ \textit{років}$$

Вибіркова дисперсія

$$\sigma^2 = \frac{\Sigma x^2}{n} - \tilde{x}^2 = \frac{11576}{30} - 17^2 = 385{,}87 - 289 = 96{,}87.$$

Скоригована дисперсія

$$S^2 = \sigma^2 \frac{n}{n-1} = 96{,}87 \frac{30}{30-1} = 100{,}21.$$

Незміщена оцінка дисперсії може бути визначена і за іншою формулою:

$$S^2 = \frac{\Sigma x^2 - n\tilde{x}^2}{n-1} = \frac{11576 - 30 \cdot 17^2}{30-1} = 100{,}21.$$

Середня помилка вибіркової середньої

$$\mu_{\tilde{x}} = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{100{,}21}{30}} = \sqrt{3{,}34} = 1{,}83 \ \textit{року}.$$

Проведемо точкову оцінку середнього стажу роботи в генеральній сукупності:

$$\tilde{x} = \bar{x} = 17 \text{ років при } \mu_{\tilde{x}} = 1{,}83 \text{ року},$$

тобто точкова оцінка генеральної середньої може бути записана так:

$$\tilde{x} \pm \mu_{\tilde{x}} = 17 \pm 1{,}83 \text{ року}.$$

Це означає, що $\tilde{x} = 17$ років є оцінкою генеральної середньої з помилкою, що дорівнює 1,83 року.

Для проведення інтервальної оцінки і побудови довірчого інтервалу визначимо граничну помилку вибіркової середньої при Р = 0,9545 і $t = 2$.

Гранична помилка вибіркової середньої

$$\varepsilon_{\tilde{x}} = t\mu_{\tilde{x}} = 2 \cdot 1{,}83 = 3{,}66 \text{ року.}$$

Побудуємо довірчий інтервал, в якому із заданим рівнем імовірності знаходиться середній стаж роботи слідчих в генеральній сукупності:

$$\tilde{x} - \varepsilon_{\tilde{x}} < \bar{x} < \tilde{x} + \varepsilon_{\tilde{x}};$$

$$17 - 3{,}66 < \bar{x} < 17 + 3{,}66;$$

$$13{,}34 < \bar{x} < 20{,}66.$$

Таким чином, довірчі межі інтервалу

$$I_{0,9545} = (13{,}34;\ 20{,}66),$$

що можна записати так: $\bar{x} = 17 \pm 3{,}66$ року.

Отже, з довірчою імовірністю $P = 0{,}9545$ можна стверджувати, що середній стаж роботи слідчих у генеральній сукупності знаходиться в інтервалі 13,34 - 20,66 року.

Точкова і інтервальна оцінка генеральної середньої в малих вибірках (n<30) проводиться аналогічно оцінці у великих вибірках лише з тією різницею, що при визначенні граничної помилки замість t - критерію нормального розподілу використовується t - критерій Стьюдента (дод. 17).

8.3. Способи формування вибіркових сукупностей

Результати вибіркового спостереження багато в чому залежать від способів формування та відбору одиниць у вибіркову сукупність. Основним принципом правильності відбору одиниць є строго об'єктивний підхід до відбору одиниць для спостереження. Дотримання цього принципу дає змогу запобігти систематичних (тенденційних) помилок і найбільш точно і повно представити генеральну сукупність.

Попередження систематичних помилок досягається в результаті застосування науково обґрунтованих способів формування вибіркової сукупності.

При формуванні вибіркової сукупності мають бути забезпечені дві умови:

1) рівні можливості для кожної одиниці генеральної сукупності потрапити у вибірку (так званий **принцип рівноможливості**);

2) досить представницька чисельність вибіркової сукупності.

В статистиці застосовуються різні види і способи формування вибіркової сукупності. В кожному конкретному випадку залежно від цілого ряду умов, а саме завдань дослідження, суті досліджуваного явища, специфіки об'єкта, обсягу сукупності, коливання ознаки, наявності матеріальних і трудових ресурсів, вибирають найбільш оптимальну систему формування вибіркової сукупності, яка визначається видом і способом відбору.

За видами розрізняють: 1) **індивідуальний відбір** - у вибірку потрапляють окремі одиниці генеральної сукупності; 2) **груповий відбір** - у вибірку потрапляють якісно однорідні групи або серії досліджуваних одиниць; 3) **комбінований відбір** як комбінація індивідуального і групового відбору.

Спосіб відбору визначає конкретний механізм або процедуру вибірки одиниць з генеральної сукупності. В практиці застосування вибіркового спостереження найбільше поширення одержали такі види вибірки: власне випадкова, механічна, типічна, серійна (гніздова) і комбінована. Ці способи можуть бути застосовані і в поєднанні один з одним.

При власне випадковій вибірці відбір одиниць з генеральної сукупності проводиться без попереднього розчленування її на будь-які групи і одиниця спостереження збігається з обліковою одиницею.

Суть випадкового відбору полягає в тому, що кожна одиниця спостереження потрапляє у вибірку випадково – за жеребом. Залежно від способу відбору одиниць розрізняють повторний і безповторний відбір.

При **повторному відборі** (за схемою поверненого шару) кожна одиниця після її реєстрації повертається до генеральної сукупності і знову може бути відібраною. Цей спосіб відбору забезпечує постійність складу генеральної сукупності. Імовірність потрапляння кожної одиниці до вибірки залишається постійною, отже, зберігається незалежність наступного витягання одиниць від попередніх.

При **безповторному відборі** (за схемою неповерненого шару) кожна одиниця після її реєстрації до генеральної сукупності не повертається і в подальшому відборі участі не бере, тобто та сама одиниця не може двічі потрапити до вибірки. Тому безповторна вибірка краще репрезентує генеральну сукупність і, отже, дає меншу помилку, ніж повторна.

На відміну від повторного відбору при безповторному відборі не зберігається постійність генеральної сукупності, а імовірність потрапляння окремих одиниць довибірки весь час змінюється (для одиниць, що залишилися, вона зростає). В зв'язку з цією особливістю до формул середньої та граничної помилок вибірки вводиться поправочний коефіцієнт, про що вже згадувалось вище.

Щоб позбутися елементів суб'єктивності при відборі одиниць з генеральної сукупності можна користуватися таблицею випадкових чисел (дод. 20)

Випадковий відбір дає добрі результати в однорідних сукупностях, тобто в тих, де варіація ознаки є незначною. Якщо ж сукупність неоднорідна і складається з різних типів явищ, то необхідно застосувати типову вибірку.

Механічний відбір – це різновидність випадкового відбору. Суть його полягає в тому, що всі одиниці генеральної сукупності розташовуються в певному порядку (за зростанням або зменшенням, за алфавітом, географічним положенням тощо), а потім суто механічно через певний інтервал відбираються одиниці у вибіркову сукупність.

Наприклад, якщо треба відібрати 100 об'єктів із генеральної сукупності чисельністю 1000 одиниць, то величина інтервалу становитиме $h{=}N : n = 1000{:}100{=}10$, тобто слід відібрати по одній одиниці з кожного десятка. Щоб забезпечити випадковість відбору доцільно першу вибірку з першого інтервалу провести за жеребом. Якщо відбір починають з третього об'єкта, то до вибірки потраплять 3-й, 13-й, 23-й і т. д. об'єкти.

Середню і граничні помилки вибірки при механічному відборі розраховують за тими самими формулами, що і для випадкового безповторного відбору, оскільки механічний відбір, як правило, проводиться безповторно.

При **типовому відборі** всю генеральну сукупність попередньо поділяють на типові групи за досліджуваною ознакою, а потім із кожної групи випадковим або механічним способом відбирають необхідну кількість одиниць. При цьому до початку відбору необхідно забезпечити принцип пропорційного представництва кожної групи відповідно з їхньою чисельністю або їхніх середніх квадратичних відхилень або дисперсій. Можливий також відбір, пропорційний обом показникам – чисельності одиниць в типових групах і ступеню варіації ознаки. Такий відбір називається **оптимальним**. На практиці найчастіше застосовують вибірку, пропорційну чисельності типових груп.

Розчленування сукупності на типові групи дає можливість усунути вплив міжгрупової (систематичної) варіації на результати вибірки, оскільки у вибірці забезпечується представництво всіх груп, що може не мати місця при випадковому відборі. Отже, самою вибіркою відтворюється (відображується) міжгрупова варіація. Залишається варіація

вибіркових даних навколо середніх (внутрішньогрупова або залишкова варіація), яку визначають по кожній із виділених типових груп.

Тому розмір середньої помилки вибірки буде визначатись тільки величиною внутрішньогрупової (залишкової) варіації ($\tilde{\sigma}^2_{в.гр}$) ознаки, яка менша від загальної на величину міжгрупової варіації. Залишкову дисперсію обчислюють як середню арифметичну зважену із середніх квадратів відхилень, які вимірюють залишкову варіацію в кожній групі.

Середня помилка типової вибірки розраховується за формулою:

$$\mu_{типол} = \sqrt{\frac{\tilde{\sigma}^2_{в.гр}}{n}}.$$

Наведемо формули граничних помилок вибірки для вибіркової середньої і частки для типової вибірки:

для середньої для частки

$$\varepsilon_{\tilde{x}} = t\sqrt{\frac{\sigma^2_{в.гр}}{n}\left(1-\frac{n}{N}\right)}; \qquad \varepsilon_{\omega} = t\sqrt{\frac{\overline{\omega(1-\omega)}}{n}\left(1-\frac{n}{N}\right)}.$$

Типовий відбір дає точніший результат порівняно з випадковим або механічним відбором, оскільки розчленуванням сукупності на типові групи забезпечується потрапляння до вибірки одиниць від усіх виділених груп і типів.

Суть **серійного відбору** полягає в тому, що відбору підлягають не окремі одиниці генеральної сукупності, а цілі серії таких одиниць. У відібраних методом випадкового безповторного або механічного відбору серіях проводять суцільний опис усіх одиниць, що до них увійшли. При цьому загальне число серій, які складають генеральну сукупність, розглядають як її загальну чисельність N_c, а кількість відібраних - як обсяг вибірки n_c.

Оскільки при серійному способі формування вибіркової сукупності кожна серія виступає як самостійна одиниця спостереження, то варіація усередині серій (внутрішньосерійна, $\sigma^2_{в.с}$) при розрахунку середньої помилки має бути виключена. Отже, середня помилка вибірки у цьому випадку залежить тільки від міжсерійної варіації $\sigma^2_{м.с}$:

$$\mu_c = \sqrt{\frac{\sigma^2_{м.с}}{n}}.$$

Формули граничних помилок вибірки для середньої і для частки при серійному відборі матимуть вигляд:

для середньої для частки

$$\varepsilon_{\tilde{x}} = t\sqrt{\frac{\tilde{\sigma}^2_{м.с}}{n_c}\left(1-\frac{n_c}{N_c}\right)}; \qquad \varepsilon_{\omega} = t\sqrt{\frac{\omega_{м.с}(1-\omega_{м.с})}{n_c}\left(1-\frac{n_c}{N_c}\right)}.$$

Комбінований відбір. Розглянуті способи вибірки на практиці застосовуються не тільки самостійно, але і в комбінуванні в різних поєднаннях і з різною послідовністю. Так, наприклад, можна комбінувати серійний відбір з власне випадковою вибіркою. При цьому генеральна сукупність спочатку поділяється на серії і відбирається необхідне число серій, а далі у відібраних серіях проводиться випадковий відбір одиниць у вибіркову сукупність. Можлива також комбінація типової і серійної вибірки, коли серії відбираються в установленому порядку з кількох типових груп.

Розрізняють також одноступінчастий і багатоступінчастий способи відбору одиниць у вибіркову сукупність.

При **одноступінчастій вибірці** кожна відібрана одиниця зразу ж підлягає вивченню. Так обстежують одиниці вибіркової сукупності при власне-випадковій вибірці.

При **багатоступінчастій вибірці** спочатку проводять відбір з генеральної сукупності окремих груп, а потім з відібраних груп формують вибірку другого, третього і т. д. порядку, яку й аналізують.

В статистичній практиці найбільш широке застосування одержали двоступінчаста і триступінчаста вибірки. Прикладом двоступінчастої вибірки може бути аналіз стану і причин злочинності в якому-небудь регіоні, при якому на першому етапі відбирають статистичні картки на злочинців за родовим об'єктом посягань, а на другому всередині кожного родового об'єкта відбираються статистичні картки за видами діянь. Прикладом **триступінчастої** вибірки є відбір домогосподарств, при якому на першій стадії з врахуванням виробничого напрямку відбирають райони, на другій – селища, і, насамкінець, на третій - окремі домогосподарства.

Багатоступінчаста вибірка дає, як правило, менш точні результати порівняно з одноступінчастою, оскільки її помилки складаються з помилок на окремих ступенях відбору. Однак на практиці, якщо одноступінчату вибірку організувати складно, використовують багатоступінчасту вибірку.

8.4. Визначення необхідної чисельності вибірки

При організації вибіркового спостереження виникає питання про те, якою повинна бути чисельність вибіркової сукупності, при якій межі

можливої помилки не перевищать деякої заздалегідь заданої дослідником величини. Необхідно встановити таку чисельність вибірки, яка з довірчим рівнем імовірності Р забезпечувала б одержання даних, що достатньо повно відображають узагальнюючі характеристики генеральної сукупності.

Надто велика вибірка приведе до нераціональних витрат трудових і матеріальних коштів, а недостатня – до великих помилок. Отже, треба встановити оптимальну чисельність вибірки, яка б гарантувала потрібну точність результатів і надійність висновків спостереження.

Необхідна чисельність вибірки залежить від таких факторів:

1. Розміру граничної помилки вибірки ε_p, тобто величини можливих відхилень показників генеральної сукупності від показників вибіркової сукупності. Чим менше розмір заданої граничної помилки, тим більшою має бути чисельність вибірки.

При визначенні необхідної чисельності вибірки гранична помилка вибірки заздалегідь задається самим дослідником залежно від характеру розв'язуємих завдань і потрібної точності висновків. На практиці звичайно виходять з того, що гранична помилка вибірки по відношенню до середньої помилки не перевищує 1-5%. Іншими словами цей процент не повинен перевищувати прийнятий довірчий рівень значущості α.

2. Ступеня варіації досліджуваної ознаки. Чим більше варіація (дисперсія, коефіцієнт варіації та ін,), тим більшою має бути чисельність вибірки.

3. Рівня довірчої імовірності Р, з яким потрібно гарантувати припустимі розміри граничної помилки вибірки. Імовірність у свою чергу пов'язана з нормованим відхиленням *t*. Чим більшим є заданий рівень довірчої імовірності Р, тим більше нормоване відхилення *t*, тим більшою має бути чисельність вибіркової сукупності.

4. Способу відбору одиниць у вибіркову сукупність (повторний або безповторний відбір).

Отже, при визначенні необхідної чисельності вибірки мають бути задані такі умови: а) розмір граничної помилки; б) рівень варіації (дисперсія, коефіцієнт варіації та ін.); в) рівень довірчої імовірності і значення нормованого відхилення, що відповідає їй.

Формули для розрахунку необхідної чисельності вибірки виводяться з формул граничних помилок для середньої і для частки шляхом відповідних алгебраїчних перетворень:

Наведемо формули необхідної чисельності вибірки для різних способів відбору:

а) **при визначенні середнього розміру ознаки**

$$n=\frac{t^2\sigma^2}{\varepsilon_{\tilde{x}}^2}$$ — власне випадкова і механічна повторна вибірка;

$$n=\frac{t^2\sigma^2 N}{\varepsilon_{\tilde{x}}^2 N+t^2\sigma^2}$$ — власне випадкова і механічна безповторна вибірка;

$$n=\frac{t^2\tilde{\sigma}^2 N}{\varepsilon_{\tilde{x}}^2 N+t^2\tilde{\sigma}^2}$$ — типова безповторна вибірка;

$$n=\frac{t^2\sigma_{м.с}^2 N_c}{\varepsilon_{\tilde{x}}^2 N_c+t^2\sigma_{м.с}^2}$$ — серійна безповторна вибірка;

б) **при визначенні частки ознаки**

$$n=\frac{t^2\omega(1-\omega)}{\varepsilon_{\omega}^2}$$ — власне випадкова і механічна повторна вибірка;

$$n=\frac{t^2\omega(1-\omega)N}{\varepsilon_{\omega}^2 N+t^2\omega(1-\omega)}$$ — власне випадкова і механічна безповторна вибірка;

$$n=\frac{t^2\overline{\omega(1-\omega)}N}{\varepsilon_{\omega}^2 N+t^2\overline{\omega(1-\omega)}}$$ — типова безповторна вибірка;

$$n=\frac{t^2\omega_{м.с}(1-\omega_{м.с})N_c}{\varepsilon_{\omega}^2 N_c+t^2\omega_{м.с}(1-\omega_{м.с})}$$ — серійна безповторна вибірка.

При обчисленні необхідної чисельності вибірки потрібно знати міру коливання досліджуваної ознаки. Однак, дисперсія ознаки або її частка p у генеральній сукупності, як правило, невідомі і визначатимуться лише після проведення вибіркового спостереження. Не знаючи цих величин, не можна визначити необхідну чисельність вибірки.

Труднощі, що виникають, можна розв'язати такими шляхами:

1. Замість фактичного значення σ_0^2 або p підставляють дані попередніх вибіркових спостережень, які проводилися в аналогічних цілях.

2. Можна провести пробні обстеження на невеликому обсязі вибірки і за даними кількох таких обстежень взяти найбільше значення дисперсії або частки.

3. Невідому величину середнього квадратичного відхилення можна знайти приблизно за величиною розмаху передбачуваної варіації ($R=x_{\max}-x_{\min}$). Доведено, що з імовірністю Р = 0,997 можна стверджувати, що розмах варіації в нормальному розподілі ознаки укладається в 6σ (крайні значення знаходяться на відстані в той або інший бік від середньої величини на 3σ), тобто R = 6σ, звідси σ = 1/6 R.

4. Якщо розрахунок необхідної чисельності вибірки проводиться для альтернативної ознаки і її частка невідома хоча б приблизно, то вона приймається рівною своєму максимальному значенню 0,5 і дає величину дисперсії, яка дорівнює 0,25 ($pq = 0{,}5 \cdot 0{,}5$).

Тоді формули для визначення необхідної чисельності вибірки при повторному і безповторному відборі набудуть відповідно такого вигляду:

$$n=\frac{0{,}25t^2}{\varepsilon_{\omega}^2} \quad \text{і} \quad n=\frac{0{,}25t^2 N}{\varepsilon_{\omega}^2 N+0{,}25t^2}.$$

Нерідко на практиці при визначенні необхідної чисельності вибірки гранична помилка вибірки задається не абсолютною величиною, а величиною відносної помилки, яка виражається в процентах. У цьому випадку і варіація ознаки має бути виражена в процентах, тобто дисперсію ознаки замінюють на коефіцієнт варіації (V).

Чисельність вибірки для випадку, коли гранична помилка задається в процентах, визначається за такими формулами:

повторний відбір безповторний відбір

$$n=\frac{t^2V^2}{(\varepsilon_p\%)^2}; \qquad n=\frac{t^2V^2N}{(\varepsilon_p\%)^2N+t^2V^2}.$$

Розглянемо приклад розрахунку необхідної чисельності вибірки при випадковому відборі.

У виправно-трудовій установі проектують вибіркове вивчення середнього строку ув'язнення засуджених злочинців. Загальна кількість засуджених становить 1000 осіб.

За проведеними раніше дослідженнями встановлено, що середній строк ув'язнення становить 7 років, а середнє коливання строку ув'язнення дорівнює 0,5 року.

Потрібно визначити, яка кількість засуджених підлягає обстеженню, щоб визначити середній строк ув'язнення з гранично допустимою помилкою, яка не перевищувала б 0,2 року. Ув'язнених відбирати способом випадкового безповторного відбору.

Надійний рівень ймовірності взяти таким, що дорівнює $P = 0{,}9545$.

За умовою задачі відомими є чисельність генеральної сукупності (N=1000 осіб), середнє коливання строку ув'язнення, виражене середнім квадратичним відхиленням $\sigma = 0,5$ року), гранична помилка вибірки ($\varepsilon_{\tilde{x}} = 0,2$ року) і рівень надійної ймовірності, якому відповідає нормоване відхилення, що дорівнює $t = 2,0$ (дод. 2).

При випадковому безповторному відборі чисельність вибірки визначають за формулою:

$$n = \frac{t^2\sigma^2 N}{t^2\sigma^2 + \varepsilon_{\tilde{x}}^2 N} = \frac{2^2 \cdot 0,5^2 \cdot 1000}{2^2 \cdot 0,5^2 + 0,2^2 \cdot 1000} = \frac{1000}{41} = 24 \text{ особи.}$$

Отже, досить вибірково обстежити 24 особи, щоб з надійною ймовірністю $P = 0,9545$ (ймовірність помилки в 5 випадках із 100) визначити середній строк ув'язнених з помилкою, яка не перевищує 0,2 роки.

Щоб мати більш високу гарантію результатів вибіркового спостереження, можна збільшити точність вибірки, тобто зменшити розміри допустимої граничної помилки.

Так, якщо граничну помилку вибірки зменшити в два рази – з 0,2 року до 0,1 року, то при тих самих умовах $\sigma = 0,5$ року і $t = 2,0$ чисельність вибірки становитиме:

$$n = \frac{t^2\sigma^2 N}{t^2\sigma^2 + \varepsilon_{\tilde{x}}^2 N} = \frac{2^2 \cdot 0,5^2 \cdot 1000}{2^2 \cdot 0,5^2 + 0,1^2 \cdot 1000} = \frac{1000}{11} = 91 \text{ особа.}$$

Таким чином, збільшення точності вибірки в два рази призводить до збільшення чисельності вибірки в 4 рази.

Для визначення необхідної чисельності вибірки за **якісною ознакою** при заданій граничній помилці можна використовувати спеціальну таблицю з уже готовими результатами (табл. 8.3.)

Як користуватися цією таблицею? Припустимо, потрібно взнати, скільки треба обстежити засуджених за нанесення умисних тяжких тілесних ушкоджень, щоб встановити серед них частку осіб в генеральній сукупності, вчинивших злочини в стані алкогольного сп'яніння з граничною помилкою вибірки, яка не перевищувала б 5%. Попередніми дослідженнями встановлено, що частка таких осіб становить 70%. Тоді на перетині рядка з величиною 70% (частка ознаки) і стовпця з величиною помилки 5% становимо, що необхідна чисельність вибірки становитиме 340 одиниць. Це означає, що при граничній помилці вибірки 5% і ймовірності Р = 0,9545 (t = 2) нам необхідно обстежити 340 осіб (статистичних карток, кримінальних справ засуджених). Якщо помилку вибірки за цих самих умов зменшити до 3%, то чисельність вибіркової сукупності становитиме вже 930 одиниць (осіб, статистичних карток, кримінальних справ), тобто в 2,7 раза більше.

Таблиця 8.3

***Чисельність спостережень, необхідна для того, щоб помилка не перевищувала заданої межі при t = 2*[1]**

При величині показника, %	Межа помилки, %					
	1	2	3	4	5	10
10	3600	900	400	230	150	37
20	6400	1600	710	400	260	65
40	9600	2400	1070	600	390	97
45	9900	2500	1100	620	400	100
55	9900	2500	1100	620	400	100
65	9100	2300	1010	570	370	92
70	8400	2100	930	530	340	85
80	6400	1600	710	400	260	65

8.5. Малі вибірки

Розглянуті вище прийоми розрахунку характеристик вибіркової сукупності (дисперсії, середньої і граничної помилок тощо) передбачають досить велику чисельність вибірки (n > 30). В той же час не завжди можливий і доцільний великий обсяг вибірки.

У практиці виробничих спостережень та в науково–дослідній роботі часто доводиться користуватися невеликими за обсягом вибірками, чисельність яких не перевищує 30 одиниць (наукові досліди, перевірка якості продукції, пов'язана зі знищенням зразків тощо). В статистиці вони дістали назву **малих вибірок.** Відповідно вибірки з чисельністю більше 30 одиниць називають **великими вибірками.**

Невеликий обсяг вибірки зменшує її точність порівняно з великою вибіркою. Проте доведено що результати, які отримані за малими вибірками, також можна поширювати на генеральну сукупність. Але тут необхідно враховувати деякі особливості, зокрема, при розрахунку середнього квадратичного відхилення. При малому обсязі вибірки слід користуватися незміщеною оцінкою дисперсії S^2.

Основи теорії малих вибірок розробив англійський математик-статистик В.Госсет (псевдонім Стьюдент). Дослідження Стьюдента по-

[1] Боярский А.Я. Таблицы для определения достоверности статистических показателей и числа наблюдений в статистическом исследовании. М., 1947.

казали, що при невеликій чисельності сукупності середнє квадратичне відхилення у вибірці значно відрізняється від середнього квадратичного відхилення в генеральній сукупності.

Оскільки середнє квадратичне відхилення генеральної сукупності є одним із параметрів кривої нормального розподілу, то використовувати функцію нормального розподілу для оцінки параметрів генеральної сукупності за даними малих вибірок в силу отримання великих помилок не правомірно.

При розрахунку середньої помилки по вибірках малої чисельності завжди треба користуватись незміщеною оцінкою дисперсії

$$S^2 = \frac{\Sigma(x_i - \tilde{x})^2}{n-1},$$

де n - 1 — **число ступенів свободи варіації** (k), під яким розуміють число одиниць, здатних приймати довільні значення, не змінюючи їх загальної характеристики (середньої).

Наприклад, проведено три спостереження: $x_1 = 4$; $x_2 = 2$; $x_3 = 6$. Середня величина

$$\bar{x} = \frac{x_1 + x_2 + x_3}{n} = \frac{4+2+6}{3} = \frac{12}{3} = 4.$$

Отже, вільно варіюючих величин залишається тільки дві, тому що третя може бути знайдена за відомими двома величинами і середньою:

$$x_3 = n\tilde{x} - (x_1 + x_2) = 3 \cdot 4 - (4 + 2) = 6;$$

$$x_1 = n\tilde{x} - (x_2 + x_3) = 3 \cdot 4 - (2 + 6) = 4 \text{ і т.д.}$$

Отже, для даного прикладу число ступенів свободи варіації дорівнює 2 ($k = n - 1 = 3 - 1 = 2$).

Стьюдент обґрунтував закон розподілу відхилень вибіркових середніх від генеральної середньої для малих вибірок. Згідно розподілу Стьюдента імовірність того, що гранична помилка не перевищить t-кратну середню помилку в малих вибірках залежить від величини t і чисельності вибірки.

Теоретичне нормоване відхилення для малих вибірок одержало назву t-критерію на відміну від t-критерію нормального розподілу, який застосовується у великих вибірках. Значення t-критерію Стьюдента наводяться в спеціальних таблицях (дод.17).

Розглянемо порядок визначення середньої і граничної помилки для малої вибірки на такому прикладі. Припустимо, для визначення питомої ваги засуджених до виправних робіт у генеральній сукупності

було відібрано 5 районів області. У розрізі районів відсоток засуджених до виправних робіт у структурі всіх засуджених становив (%); 0,6; 0,2; 0,8; 0,4; 0,5.

Середній відсоток засуджених до виправних робіт становитиме:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{2{,}5}{5} = 0{,}5\ \%.$$

Судячи з окремих спостережень, величина відсотка засуджених сильно варіює і середня лише по п'яти спостереженнях може мати велику помилку.

Для розрахунку помилок вибірки визначимо незміщену оцінку дисперсії

$$S^2 = \frac{\Sigma x^2 - n\tilde{x}^2}{n-1} = \frac{1{,}45 - 5 \cdot 0{,}5^2}{5-1} = \frac{1{,}45 - 1{,}25}{4} = \frac{0{,}20}{4} = 0{,}05.$$

Звідки $S = \sqrt{0{,}05} = 0{,}2236\ \%$.

Розрахуємо середню помилку вибіркової середньої, де замість середнього квадратичного відхилення використовується його незміщена оцінка:

$$\mu_{\tilde{x}} = \frac{S}{\sqrt{n}} = \frac{0{,}2236}{\sqrt{5}} = \frac{0{,}2236}{2{,}236} = 0{,}10\ \%.$$

За таблицями Стьюдента (дод. 17) встановимо, що при довірчій імовірності $P = 0{,}95$ (рівень значущості $\alpha = 0{,}05$) і при $k = n - 1 = 5 - 1 =$ 4 ступенях свободи варіації $t = 2{,}78$. Тоді гранична помилка вибірки дорівнює

$$\varepsilon_{\tilde{x}} = t\mu_{\tilde{x}} = 2{,}78 \cdot 0{,}10 \simeq 0{,}28\%.$$

Отже, з імовірністю $P = 0{,}95$ можна стверджувати, що величина відсотка засуджених до виправних робіт у генеральній сукупності становитиме $0{,}5 \pm 0{,}28$ %, або від 0,22 до 0,78 %.

Як бачимо з прикладу, межі випадкових коливань при малих вибірках досить великі і можуть бути скорочені за рахунок збільшення чисельності вибірки і зменшення коливання (дисперсії) ознаки.

Якщо б ми використали для розрахунку довірчих меж генеральної середньої таблицю інтегралу імовірностей (дод. 16), то t було б рівним 1,96 і $\varepsilon_{\tilde{x}} = t\mu_{\tilde{x}} = 1{,}96 \cdot 0{,}10 = 0{,}20$ %, тобто довірчий інтервал був би вужчим (від 0,30 до 0,70 %).

Малі вибірки в силу своєї невеликої чисельності навіть при найретельнішій організації спостереження не відображають достатньо точно показники генеральної сукупності. Тому результати малих вибірок рідко використовуються для встановлення надійних меж, в яких знаходяться характеристики генеральної сукупності.

Критерій Стьюдента застосовується головним чином для перевірки статистичних гіпотез щодо істотності відмінностей між показниками двох або кількох малих вибірок.

8.6. Дисперсійний аналіз

Дисперсійний аналіз – це метод статистичної оцінки надійності проявлення залежності результативної ознаки від одного або кількох факторів. За допомогою методу дисперсійного аналізу проводиться перевірка статистичних гіпотез відносно середніх в кількох генеральних сукупностях, які мають нормальний розподіл.

Суть цього методу полягає в статистичному вивченні вірогідності впливу одного або кількох факторів, а також їх взаємодії на результативну ознаку. Відповідно до цього за допомогою дисперсійного аналізу вирішуються три основних завдання:

1) дати загальну оцінку істотності відмінностей між груповими середніми;

2) оцінити вірогідність взаємодії факторів;

3) оцінка істотності відмінностей між парами середніх. Найчастіше такі завдання доводиться вирішувати дослідникам для вивчення впливу кількох факторів на результативну ознаку одночасно

Розв'язання задач дисперсійного аналізу ґрунтується на законі розкладання (додавання) варіації, відповідно до якого загальна варіація (коливання) результативної ознаки поділяється на дві: варіацію, зумовлену дією досліджуваного фактора (факторів), і варіацію, зумовлену дією випадкових причин, тобто

$$\sigma^2_{заг} = \sigma^2_{м.гр} + \sigma^2_{в.гр}.$$

Припустимо, що досліджувана сукупність поділена за факторною ознакою на кілька груп, кожна з яких характеризується своєю середньою величиною результативної ознаки. Варіацію цих величин можна пояснити двома видами причин: тими, що діють на результативну ознаку систематично і піддаються регулюванню в ході здійснюваного експерименту і тими, які регулюванню не піддаються. Очевидно, що міжгрупова (факторна або систематична) варіація залежить переважно від дії досліджуваного фактора, а внутрішньогрупова (залишкова або випадкова) – від дії випадкових факторів.

Щоб оцінити надійність відмінностей між груповими середніми, потрібно визначити міжгрупову та внутрішньогрупову варіації. Якщо міжгрупова (факторна) варіація значно перевищує внутрішньогрупову (залишкову) варіацію, то фактор впливав на результативну ознаку, істотно змінюючи значення групових середніх величин. Але виникає питання, яке співвідношення між міжгруповою і внутрішньогруповою варіаціями можна розглядати як достатнє для висновку про вірогідність (істотність) відмінностей між груповими середніми.

Щоб оцінити істотність відмінностей між середніми і сформулювати висновки з перевірки нульової гіпотези ($H_o : \bar{x}_1 = \bar{x}_2 = ... = \bar{x}_n$) у дисперсійному аналізі використовується своєрідний норматив – F-критерій, закон розподілу якого встановив Р.Фішер. Цей критерій являє собою відношення двох дисперсій: факторної, породжуваної дією досліджуваного фактора, та залишкової, зумовленої дією випадкових причин:

$$F = \frac{S_1^2}{S_2^2}, \text{ де } S_1^2 > S_2^2.$$

Дисперсійне відношення F = $S_1^2 : S_2^2$ американським статистиком Снедекором запропоновано позначати літерою F на честь винахідника дисперсійного аналізу Р.Фішера.

Якщо в експерименті перевіряють вплив кількох факторів (A, B, C і т. д.) на результативну ознаку одночасно, то дисперсія, що обумовлена дією кожного з них, має бути порівняна з $S^2_{в.гр}$, тобто

$$F_{факт}(A) = \frac{S_A^2}{S^2_{в.гр}}; \; F_{факт}(B) = \frac{S_B^2}{S^2_{в.гр}}; \; F_{факт}(C) = \frac{S_C^2}{S^2_{в.гр}} \quad \text{і т. д.}$$

Якщо значення факторної дисперсії значно більше залишкової, то фактор істотно впливав на результативну ознаку і навпаки.

У багатофакторних експериментах крім варіації, зумовленої дією кожного фактора, практично завжди є варіація, яка зумовлена взаємодією факторів $\left(S^2_{AB},\ S^2_{AC},\ S^2_{BC},\ S^2_{ABC}\right)$. Суть взаємодії полягає в тому, що ефект одного фактора істотно змінюється на різних рівнях другого.

Взаємодія факторів також має бути оцінена шляхом порівняння відповідних дисперсій з $S^2_{в.гр}$:

$$F_{факт}(AB) = \frac{S^2_{AB}}{S^2_{в.гр}}; \qquad F_{факт}(AC) = \frac{S^2_{AC}}{S^2_{в.гр}};$$

$$F_{факт}(BC)=\frac{S^2_{BC}}{S^2_{в.гр}};\qquad F_{факт}(ABC)=\frac{S^2_{ABC}}{S^2_{в.гр}}$$

Обчислюючи фактичне значення критерію F, у чисельнику береть більшу із дисперсій, тому F > 1. Очевидно, що чим більше критерій F, тим значнішим є розбіжності між дисперсіями. Якщо F = 1, то питання про оцінку істотності відмінностей дисперсій знімають.

Для визначення меж випадкових коливань відношення дисперсій Р. Фішер розробив спеціальні таблиці F-розподілу (дод. 18 і 19). Критерій F функціонально зв'язаний з імовірністю і залежить від числа ступенів свободи варіації k_1 і k_2 двох порівнюваних дисперсій. Звичайно використовуються дві таблиці, що дозволяють робити висновки про гранично високе значення критерію F для рівнів значущості 0,05 і 0,01. Рівень значущості 0,05 (або 5%) означає, що тільки в 5 випадках із 100 критерій F може приймати значення, що дорівнює вказаному в таблиці або вище його. Зниження рівня значущості з 0,05 до 0,01 призводить до збільшення значення критерію F між двома дисперсіями в силу дії тільки випадкових причин.

Значення критерію F також залежить безпосередньо від числа ступенів свободи двох порівнюваних дисперсій. Якщо число ступенів свободи прямує до нескінченості ($k\to\infty$), то відношення F для двох дисперсій прямує до одиниці.

Табличне значення критерію F показує можливу випадкову величину відношення двох дисперсій при заданому рівні значущості і відповідному числі ступенів свободи для кожної з порівнюваних дисперсій. В зазначених таблицях наводиться величина F для вибірок, зроблених з однієї і тієї самої генеральної сукупності, де причини зміни величин тільки випадкові.

Розглянемо порядок визначення числа ступенів свободи в дисперсійному аналізі. Число ступенів свободи, що відповідає загальній сумі квадратів відхилень $(W_{заг}=W_{м.гр}+W_{в.гр})$, розкладається на відповідні компоненти аналогічно розкладанню сум квадратів відхилень , тобто загальне число ступенів свободи (k_o) розкладається на число ступенів свободи для міжгрупової (k_1) і внутрішньогрупової (k_2) варіацій.

Так, якщо вибіркова сукупність, що складається з N спостережень, поділена на **m** груп (число варіантів досліду) і **n** підгруп (кількість повторностей), то число ступенів свободи k відповідно становитиме:

а) для загальної суми квадратів відхилень $(W_{заг})$

$$k_o = m\cdot n - 1 = N - 1;$$

б) для міжгрупової суми квадратів відхилень $(W_{м.гр})$

$$k_1 = m - 1;$$

в) для внутрішньогрупової суми квадратів відхилень $(W_{в.гр})$

$$k_2 = k_o - k_1 = (N-1) - (m-1) = N - m,$$

або

$$k_2 = m\,(n-1).$$

Згідно з правилом додавання варіації:

$$k_o = k_1 + k_2;\ N-1 = (m-1) + (N-1) - (m-1) = N-1.$$

Наприклад, якщо сформовано чотири варіанти досліду (m = 4) у п'яти повторностях кожен (n = 5), і загальна кількість спостережень N= = m · n = 4 · 5 = 20, то число ступенів свободи відповідно дорівнює:

$$k_o = N-1 = 20-1 = 19;\quad k_1 = m-1 = 4-1 = 3;$$

$$k_2 = k_o - k_1 = 19-3 = 16;\ \text{або}\ k_2 = N-m = 20-4 = 16;$$

або

$$k_2 = m\,(n-1) = 4\,(5-1) = 16.$$

Перевіримо розрахунок: $k_o = k_1 + k_2$; 19 = 3 + 16.

Знаючи суми квадратів відхилень і число ступенів свободи, можна визначити незміщені (скориговані) оцінки для трьох дисперсій:

$$S^2_{заг}=\frac{W_o}{k_o}=\frac{W_o}{N-1};\quad S^2_{м.гр}=\frac{W_{м.гр}}{k_1}=\frac{W_{м.гр}}{m-1};$$

$$S^2_{в.гр}=\frac{W_{в.гр}}{k_2}=\frac{W_{в.гр}}{N-m}.$$

Нульову гіпотезу H_0 за критерієм F перевіряють так само, як і за t-критерієм Стьюдента. Щоб прийняти рішення з перевірки H_0, необхідно розрахувати фактичне значення критерію $F_{факт}$ і порівняти його з табличним значенням F_α для прийнятого рівня значущості α і відповідного числа ступенів свободи k_1 і k_2 для двох дисперсій.

Значення F_a знаходять за таблицями (дод 18 і 19) на перетині відповідного стовпця (число ступенів свободи для більшої дисперсії k_1) і рядка (число ступенів свободи для меншої дисперсії k_2). Якщо $F_{факт}>F_\alpha$, то згідно з прийнятим рівнем значущості можна зробити висновок, що відмінності вибіркових дисперсій визначаються не лише випадковими факторами; вони істотні. Нульову гіпотезу в цьому випадку відхиляють і є підстава стверджувати, що фактор істотно впливає на результативну ознаку. Якщо ж $F_{факт}<F_\alpha$, то нульову гіпотезу приймають і є підстава стверджувати, що відмінності між порівнюваними дисперсіями знаходяться в

межах можливих випадкових коливань: дія фактора на результативну ознаку не є істотною.

Дисперсійний аналіз включає чотири послідовно виконувані етапи: 1) визначення і розкладання варіації; 2) визначення числа ступенів свободи варіації; 3) визначення дисперсій; 4) аналіз дисперсій і прийняття рішення з перевірки нульової гіпотези.

Найбільш трудомісткою частиною дисперсійного аналізу є перший етап – визначення і розкладання варіації за джерелом її утворення. Методику розкладання загального обсягу варіації розглянуто в розділі 7.3.

Порядок проведення дисперсійного аналізу при групуванні даних за однією ознакою розглянемо на прикладі вивчення впливу рівня безробіття на коефіцієнт (рівень) злочинності (див. розділ 7.3). З цією метою було розглянуто порядок визначення і розкладання загального обсягу варіації коефіцієнта злочинності на варіацію, пов'язану з дією досліджуваного фактора (рівня безробіття), – міжгрупову і варіацію, зумовлену іншими неврахованими факторами, – внутрішньогрупову, або залишкову.

На підставі вихідних даних прикладу розділу 7.3. і результатів розкладання загального обсягу варіації на основні частини методом дисперсійного аналізу перевіримо статистичну гіпотезу відносно середніх у генеральних сукупностях.

Сформулюємо нульову та альтернативну гіпотези:

$$H_0 : \bar{x}_1 = \bar{x}_2 = \bar{x}_3 ; \qquad H_x : \bar{x}_1 \neq \bar{x}_2 \neq \bar{x}_3 .$$

Рівень значущості візьмемо таким, що дорівнює $\alpha = 0{,}05$.

Найбільш потужним критерієм перевірки H_0 є F – критерій Фішера.

Для перевірки H_0 і формулювання висновків за результатами дисперсійного аналізу потрібно встановити фактичне значення F – критерію Фішера і порівняти його з табличним значенням (F_α).

З метою обчислення фактичного значення F – критерію виконаємо потрібні операції відповідно до етапів дисперсійного аналізу.

Перший етап дисперсійного аналізу – визначення і розкладання варіації – дав такі результати:

$$W_{заг} = W_{м.гр} + W_{в.гр} ; \qquad 54{,}50=50{,}92+3{,}58;$$
$$100\%=93{,}4\%+6{,}6\%.$$

Отже, 93,4% загальної варіації коефіцієнта злочинності припадає на частку досліджуваного фактора (рівня безробіття), а 6,6% варіації зумовлено випадковими факторами.

Другий етап дисперсійного аналізу – визначення числа ступенів свободи варіації. Встановимо число ступенів свободи варіації для кожної суми квадратів відхилень ($W_{заг}$, $W_{м.гр}$, $W_{в.гр}$). Для нашої задачі загальна кількість спостережень N = 12, кількість варіантів досліду m = 3, кількість повторностей n = 4.

Тоді число ступенів свободи варіації для загальної дисперсії:

$$k_0 = N - 1 = 12 - 1 = 11;$$

для міжгрупової

$$k_1 = m - 1 = 3 - 1 = 2;$$

для внутрішньогрупової визначають за різницею, як і саму залишкову варіацію:

$$k_2 = k_0 - k_1 = (N - 1) - (m - 1) = 11 - 2 = 9,$$

або за іншими формулами:

$$k_2 = m\,(n - 1) = 3\,(4 - 1) = 3 \cdot 3 = 9;$$
$$k_2 = N - m = 12 - 3 = 9.$$

Перевіримо розрахунки: $k_0 = k_1 + k_2$; $11 = 2 + 9$.

Третій етап дисперсійного аналізу – визначення дисперсій.

Для дисперсійного аналізу становить інтерес міжгрупова і внутрішньогрупова дисперсії; загальна дисперсія в аналізі не бере участі, тому її не обчислюємо:

$$S^2_{м.гр} = \frac{W_{м.гр}}{k_1} = \frac{50{,}92}{2} = 25{,}46 ; \quad S^2_{в.гр} = \frac{W_{в.гр}}{k_2} = \frac{3{,}58}{9} = 0{,}40.$$

Четвертий етап дисперсійного аналізу – аналіз дисперсій і формулювання висновків щодо перевірки нульової гіпотези.

Порівняємо дисперсії, тобто знайдемо фактичне значення F – критерію Фішера:

$$F_{факт} = \frac{S_1^2}{S_2^2} = \frac{S^2_{м.гр}}{S^2_{в.гр}} = \frac{25{,}46}{0{,}40} = 63{,}65.$$

Для перевірки H_0 потрібно визначити табличне значення F – критерію Фішера (дод.18) і порівняти його із знайденим фактичним значенням.

Більшій дисперсії $S^2_{м.гр}$ відповідає число ступенів свободи варіації k_1=2 (чисельник відношення), меншій $S^2_{в.гр}$ - число ступенів свободи варіації k_2 = 9 (знаменник відношення). Отже, згідно з дод. 18 теоретичне (табличне) значення F – критерію перебуває на перетині другого стовпця і дев'ятого рядка: $F_{0,05}$ = 4,26.

Якщо підвищити рівень значущості до α = 0,01, то при тих самих ступенях свободи варіації $F_{0,01}$ становитиме 8,02 (дод. 19).

Результати розрахунків оформимо в табл. 8.4.

Таблиця 8.4

Дані для аналізу дисперсій

Варіація	Обсяг варіації	Ступінь свободи варіації	Дисперсія	Значення F-критерію	
	W_i	k_i	S_i^2	$F_{факт}$	$F_{0,05}$
Міжгрупова (рівень безробіття)	50,92	2	25,46	63,65	4,26
Залишкова (випадкова)	3,58	9	0,40	1	-
Загальна	54,50	11	4,95	-	-

Порівняємо фактичне і табличне значення F – критерію:

$$F_{факт} > F_{табл}; \quad 63,65 > 4,26; \quad 63,65 > 8,02.$$

Оскільки $F_{факт} > F_{табл}$, висунуту нульову гіпотезу про випадкові розбіжності в групових середніх потрібно відхилити і прийняти альтернативну гіпотезу: значення генеральних середніх істотно розрізняються. Іншими словами, фактичні дані не узгоджуються з нульовою гіпотезою. Отже, вплив рівня безробіття на коефіцієнт злочинності є вірогідним і істотним.

Критерій F дозволяє встановити наявність або відсутність істотних зв’язків між груповими середніми в цілому, однак він не показує, між якими середніми різниця істотна, а між якими неістотна.

Тому, якщо проведений дисперсійний аналіз призвів до відмови від нульової гіпотези, що передбачає рівність середніх, та показав істотність впливу перевіряємого фактора на результативну ознаку, то цей загальний висновок може бути конкретизований, доповнений перевіркою істотності відмінностей між парами середніх.

Для перевірки нульової гіпотези про вірогідність відмінностей між парами середніх використовується ряд методик і критеріїв.

При нерівних чисельностях вибірок доцільно застосовувати метод контрастів Шеффе. У разі невеликої кількості варіантів досліду (не більше трьох) можливе використання t-критерію Стьюдента. При багатоваріантному досліді (більше трьох) застосування цього критерію дає збільшену кількість вірогідних зв’язків.

Методика застосування Q-критерію Тьюкі така:

1) будують ранжирований ряд групових середніх;

2) визначають попарні різниці між середніми, спочатку між середніми, що стоять поряд, потім між середніми, віддаленими одна від одної на дві, три позиції і т. д.;

3) обчислюють фактичне значення Q-критерію Тьюкі як відношення знайдених різниць до середньої помилки вибірки:

$$Q_{факт} = \frac{\left|\tilde{x}_i - \tilde{x}_j\right|}{\sqrt{\frac{S^2_{зал}}{n}}},$$

де $S^2_{зал}$ – дисперсія, обчислена у дисперсійному аналізі, n – кількість спостережень у кожному варіанті досліду (кількість повторностей);

4) визначають табличне значення Q-критерію Тьюкі (дод. 21); при цьому значення Q_α різні для середніх, віддалених одна від одної на одну, дві, три позиції і т. д. і залежать від рівня значущості (його треба взяти таким самим, як і в дисперсійному аналізі), числа ступенів свободи і порядку різниць l ($l = 2$ для сусідніх в ранжированому ряду середніх, $l = 3$ для середніх, які відстоять одна від одної на дві позиції і т.д.);

5) зіставляють $Q_{факт}$ і Q_α; якщо $Q_{факт} > Q_\alpha$, то нульову гіпотезу про рівність двох середніх відхиляють, а різницю між середніми вважають істотною (вірогідною), якщо ж $Q_{факт} < Q_\alpha$, то нульову гіпотезу приймають, а різницю між середніми визнають неістотною оскільки, вона знаходиться в межах можливих випадкових коливань.

Вірогідність різниці між парами середніх може бути оцінена також шляхом співставлення її з гранично можливою помилкою вибірки ε_p, яка вказує межі граничних випадкових коливань і дістала назву **найменшої істотної різниці (НІР).** Можлива гранична помилка при використанні t-критерію Стьюдента розраховується за формулою $\varepsilon_p = t\,\bar{\mu}_{1-2}$, при використанні більш строгого Q-критерію Тьюкі – за формулою $\varepsilon_p = Q\,\bar{\mu}_{1-2}$.

Якщо різниця між двома порівнюваними середніми, більша за абсолютною величиною, ніж можлива гранична помилка (НІР), то роблять висновок про істотність різниці цих двох середніх. Якщо ж можлива гранична помилка буде більшою від фактичної різниці, то різниця між двома середніми лежить у межах можливих випадкових коливань, тобто вона є невірогідною.

Проведемо оцінку істотності різниці між парами середніх по розглянутому прикладу оскільки був одержаний загальний висновок щодо вірогідності впливу рівня безробіття на коефіцієнт злочинності.

Нагадаємо, що за даними аналізованого досліду вже обчислено три групові середні (табл. 7.3) $\tilde{x}_1 = 9{,}5$; $\tilde{x}_2 = 12{,}0$; $\tilde{x}_3 = 15{,}0$; (злочинів).

Залишкова дисперсія, визначена в дисперсійному аналізі, становила $S^2_{зал} = 0{,}40$, їй відповідало число ступенів свободи варіації $k_2 = 9$; кількість повторностей $n = 4$.

Сформулюємо H_0 і H_α. За даними досліду можливі такі три порівняння:

$$H_0 : \bar{x}_1 = \bar{x}_2;\ \ H_a : \bar{x}_1 \neq \bar{x}_2;$$

$$H_0 : \bar{x}_2 = \bar{x}_3;\ \ H_a : \bar{x}_2 \neq \bar{x}_3;$$

$$H_0 : \bar{x}_1 = \bar{x}_3;\ \ H_a : \bar{x}_1 \neq \bar{x}_3.$$

Рівень значущості візьмемо таким, що дорівнює $\alpha = 0{,}05$.

Найбільш потужним критерієм перевірки H_0 такого роду є Q – критерій Тьюкі.

Для перевірки H_0 потрібно обчислити фактичні значення критерію $Q_{факт}$ і порівняти їх з табличним значенням $Q_{0,05}$.

Побудуємо ранжирований ряд групових середніх за величиною середнього коефіцієнта злочинності (злочинів):

$$\tilde{x}_1 = 9{,}5;\ \tilde{x}_2 = 12{,}0;\ \tilde{x}_3 = 15{,}0.$$

Визначимо середню помилку вибірок:

$$\bar{\mu}_{1-2} = \sqrt{\frac{S^2_{зал}}{n}} = \sqrt{\frac{0{,}40}{4}} = \sqrt{0{,}10} = 0{,}316 \text{ злочина.}$$

Обчислимо фактичні значення Q – критерію Тьюкі для кожної пари середніх:

$$Q_1 = \frac{|\tilde{x}_1 - \tilde{x}_2|}{\bar{\mu}_{1-2}} = \frac{|9{,}5 - 12{,}0|}{0{,}316} = \frac{9{,}5}{0{,}316} = 7{,}91;$$

$$Q_2 = \frac{|\tilde{x}_2 - \tilde{x}_3|}{\bar{\mu}_{1-2}} = \frac{|12{,}0 - 15{,}0|}{0{,}316} = \frac{3{,}0}{0{,}316} = 9{,}49;$$

$$Q_3 = \frac{|\tilde{x}_3 - \tilde{x}_1|}{\bar{\mu}_{1-2}} = \frac{|15{,}0 - 9{,}5|}{0{,}316} = \frac{5{,}5}{0{,}316} = 17{,}40.$$

За таблицею "Значення Q – критерію Тьюкі (дод.21) встановимо критичні значення критерію Q при $\alpha = 0{,}05$; $k = 9$:

для різниць першого ($l = 2$) порядку $Q_{0,05} = 3{,}199$;

для різниць другого ($l = 3$) порядку $Q_{0,05} = 3{,}949$.

Порівняємо фактичні і табличні значення критерію (табл. 8.5).

Таблиця 8.5

Дані для аналізу статистичної оцінки різниць між парами середніх

Фактична різниця між середніми, злочинів	Значення Q – критерію Тьюкі		Можлива гранична помилка	
	$Q_{факт}$	$Q_{0,05}$	?$_{факт}$	?$_{0,05}$
Першого порядку				
$\tilde{x}_2 - \tilde{x}_1 = 2{,}5$	7,91	3,199	2,5	1,01
$\tilde{x}_3 - \tilde{x}_2 = 3{,}0$	9,49	3,199	3,0	1,01
Другого порядку				
$\tilde{x}_3 - \tilde{x}_1 = 5{,}5$	17,40	3,949	5,5	1,25

Оскільки для всіх трьох порівнянь $Q_{факт} > Q_{0,05}$, нульові гіпотези про рівність середніх у генеральних сукупностях відхиляють.

Ті самі висновки матимемо, порівнюючи можливу граничну помилку (НІР) з фактичною різницею між парами середніх.

Визначимо ε_p:

а) для різниць першого порядку

$$\varepsilon_p = Q_{0,05} \cdot \mu_{1-2} = 3{,}199 \cdot 0{,}316 = 1{,}01 \text{ злочина;}$$

б) для різниць другого порядку

$$\varepsilon_p = Q_{0,05} \cdot \mu_{1-2} = 3{,}949 \cdot 0{,}316 = 1{,}25 \text{ злочина.}$$

Можлива гранична помилка (НІР) показує, що внаслідок випадкового варіювання різниця між парами середніх першого порядку може досягти 1,01 злочина, другого порядку – 1,25 злочина.

Порівняння фактичних і гранично можливих значень різниць між парами середніх (табл. 8.5) показує, що всі три фактичні різниці значно перевищують межі можливих випадкових коливань.

Отже, з імовірністю помилки тільки в 5 випадках із 100 можна стверджувати, що різні рівні безробіття істотно впливають на коефіцієнт (рівень) злочинності.

Залежно від кількості факторів, що визначають варіацію результативної ознаки, дисперсійний аналіз поділяють на однофакторний і багатофакторний.

За аналогією з комбінаційними групуваннями багатофакторні моделі дисперсійного аналізу мають незаперечну перевагу порівняно з однофакторними моделями: вони дають змогу виявити ступінь впливу не тільки кожного фактора окремо, а й їхню взаємодію. Це завдання розв'язується за допомогою побудови комбінаційних групувань і таблиць. Відмінність багатофакторного аналізу від однофакторного полягає в тому, що загальний обсяг варіації розкладають на більшу кількість компонентів. У міру поділу сукупності на групи і підгрупи ускладнюються розрахунки з розкладання загального обсягу варіації на складові частини, а також аналіз дисперсій.

Схема розкладання загальної варіації залежить від формування груп. Воно може бути випадковим (спостереження однієї групи не пов'язане із спостереженнями другої групи) і невипадковим (спостереження двох вибірок зв'язані між собою спільністю умов експерименту). Відповідно дістають незалежні і залежні вибірки.

На практиці в більшості випадків доводиться розглядати залежні вибірки, коли умови для груп і підгруп вирівнюються. Так, у наукових дослідженнях всю сукупність спостереждень розбивають на блоки, з максимально вирівняними умовами. При цьому кожен варіант досліду отримує рівні можливості бути представленим у всіх блоках, чим досягається вирівнювання умов для всіх перевірюваних варіантів досліду. Такий метод побудови досліду дістав назву методу **рендомізованих блоків**.

Розглянемо схему розкладання загального обсягу варіації на складові частини для випадку з двома факторами (А і В). Джерелами варіації при групуванні даних за двома ознаками є: перший фактор – А, другий фактор – В, їхня взаємодія – АВ, залишкове варіювання.

Тоді загальну суму квадратів відхилень можна подати, як

$$W_0 = W_A + W_B + W_{AB} + W_{в.гр}.$$

Розклад загального обсягу варіації доцільно здійснити в два етапи. Для цього потрібно побудувати дві таблиці. На першому етапі із загального варіювання слід виділити варіацію, пов'язану з двома факторами, і залишкову варіацію:

$$W_0 = W_{A+B} + W_{в.гр},$$

а на другому етапі – відповідно

$$W_{A+B} = W_A + W_B + W_{AB}.$$

Такий порядок розкладання загального обсягу варіації справедливий для незалежних вибірок. Якщо ж вибірки залежні (наприклад, рендомізовані блоки, латинський квадрат тощо), то з'являється новий компонент варіації, зв'язаний з повтореннями.

Тоді схема розкладання загальної варіації набирає такого вигляду:

І етап:

$$W_0 = W_{A+B} + W_{повт} + W_{в.гр};$$

ІІ етап:

$$W_{A+B} = W_A + W_B + W_{AB}.$$

$$W_0 = W_A + W_B + W_{AB} + W_{повт} + W_{в.гр}.$$

Для випадку з трьома факторами і залежними вибірками схема розкладання загальної варіації ускладнюється:

$$W_0 = W_A + W_B + W_C + W_{AB} + W_{AC} + W_{BC} + W_{ABC} + W_{повт} + W_{в.гр}.$$

Факторні моделі з великою кількістю факторів (трьома і більше), доцільно досліджувати кореляційним методом з використанням ЕОМ (див. розд 9).

Питання для самоконтролю

1. У чому суть вибіркового спостереження, необхідність і доцільність його застосування?
2. У чому полягають переваги вибіркового спостереження порівняно з суцільним?
3. Які наукові умови застосування вибіркового спостереження?
4. Що означає репрезентативність вибірки?
5. Які є помилки вибірки?
6. Охарактеризуйте середню помилку вибірки, наведіть алгоритм її розрахунку.
7. Охарактеризуйте граничну помилку вибірки, наведіть алгоритм її розрахунку.
8. Що таке довірчий рівень імовірності і рівень значущості?
9. Які Ви знаєте способи відбору одиниць у вибіркову сукупність?
10. У чому полягає суть випадкового повторного і безповторного відбору?
11. У чому полягає суть механічного і типового відбору?
12. У чому полягає суть серійного і комбінованого відбору?
13. Як розраховується необхідна чисельність вибірки? Від яких факторів залежить чисельність вибірки?

14. У чому полягають особливості розрахунку необхідної чисельності вибірки при випадковому, типовому і серійному відборі?

15. Що таке точкова і інтервальна оцінка генеральної середньої, довірчий інтервал?

16. Що таке мала вибірка? У чому полягають особливості визначення помилок вибірки в малих вибірках?

17. Для визначення середнього віку студентів факультету проведена випадкова безповторна вибірка 100 студентів із загального числа 1000 чол.

У результаті обстеження отримані такі дані:

Вік, років	До 18	19	20	21	22	Понад 22
Число студентів, чол.	25	17	22	18	10	8

Визначте: а) середній вік студентів за вибіркою; б) середню і граничну помилку вибірки з імовірністю 0,954; в) з тією самою імовірністю довірчий інтервал для середньої і частки студентів, вік яких становить 21 рік і більше.

18. У чому полягає суть і призначення дисперсійного аналізу?

19. Наведіть принципову схему дисперсійного аналізу.

20. Які критерії застосовуються в дисперсійному аналізі для оцінки різниць кількох середніх?

21. Розкрийте смисл нульової гіпотези в дисперсійному аналізі.

22. Що показує фактичне і табличне значення F- критерію Фішера?

23. Наведіть і поясніть суть формули фактичного дисперсійного відношення

24. Наведіть алгоритм дисперсійного аналізу при групуванні даних за однією ознакою.

25. За яких обставин і чому проводиться попарне порівняння середніх?

26. Що таке найменша істотна різниця (НІР)